Creation of Fiat Boundaries in Higher order Phenomenon

Omair Chaudhry & William Mackaness
Institute of Geography
School of GeoSciences,
University of Edinburgh,
Drummond St,
Edinburgh EH8 9XP
O.Chaudhry@sms.ed.ac.uk
William.mackaness@ed.ac.uk

Keywords: Fiat and Bona fide boundaries, gravitational modelling, cluster analysis, spatial data mining

Abstract

Spatial database generalisation involves creation of higher order objects such as cities, forest regions, and mountain ranges from lower order objects in the source database (such as buildings, trees and individual groups of hills). In order to create these higher order objects in the database it is necessary to identify their extent in terms of their boundaries. These boundaries can then be used to determine the instances of source objects in terms of their higher order target objects. This paper presents an approach for generation of indeterminate boundaries for objects at higher levels of abstraction (1:250,000) from a source database at a notional scale of 1:1250/1:10,000 (The MasterMap database of the National Mapping Agency of Great Britain, the Ordnance Survey (OS). The paper illustrates the importance of modelling density and clustering for the identification of these boundaries. The results along with an overview of the implementation are presented.

1.0 Introduction

The boundaries to geographic phenomena often lack crispness (Smith and Mark 2003; Winter and Thomas 2002). Yet in both spatial analysis and cartographic display, it is often necessary to define boundaries in the form of a simple vector. Thus a bounding polygon might define the extent of a city, a mountainous region or an area of outstanding natural beauty, but these are not 'lines' we find in reality. Traditionally drawing such boundaries was the preserve of the cartographer, who brought to bear their cartographic and geographic knowledge in deciding an appropriate line, whilst poring over aerial photographs. In the context of GIS, perceived wisdom is that we record phenomena at the finest level of detail, and via the process of model and cartographic generalisation, derive 'higher order' phenomenon from that fine detail. Thus we might use group theory or density measures applied to buildings in order to define 'city', or groups of tree stands in order to define 'forested regions', or hypsometric analysis to group hills into 'mountain chains'. These higher order phenomena do not exist explicitly in the database – they must be made explicit by applying various spatial analysis techniques to phenomena at the fine scale.

This research presents a spatial data mining approach to the derivation of boundaries of higher order phenomena usually represented by objects at 1:250,000 from OS MasterMap (1:1250/1:10,000). Spatial data mining involves extraction of implicit knowledge and spatial relations which are not explicitly represented in spatial databases (Wang W., Yang J., and Muntz 1997). Clustering is one of the main techniques used in spatial data mining for knowledge discovery (Miller and Han 2001). Spatial clustering algorithms exploit spatial relationships among data objects in order to determine inherent grouping of the input data. In this paper we propose a clustering algorithm applied to source objects (buildings and area patches) in the source database (OS MasterMap 1:1250/1:10,000) in order to find the extent of a city or forest/lake at higher levels of abstraction.

The paper is organized as follows: Section 2 gives an overview on boundaries in geographic phenomena and outlines the limitation of proximity based clustering; Section 3 gives an overview of the proposed methodology; Section 4 presents the design, implementation and the results obtained;

Section 5 outlines a few utilities of the results obtained. The paper concludes with thoughts on further research.

2.0 Boundaries in Geographic Domain

Boundaries of natural geographic phenomena are much less distinct and discrete than say, the typical boundary of table top objects. This observation is reflected in research on the modelling of fuzzy boundaries. Nevertheless a boundary separates the entity from its environment and is one of the marks of its individuality (Casati, Smith, and Varzi 1998). Geographic boundaries are dependent upon the scale of observation; sometime they are vague and fuzzy (for instance the boundary of a Forest, or mountainous region). Boundaries can be divided into two basic types: bona-fide boundaries and fiat-boundaries (Smith and Varzi 2000; Smith and Varzi 1997). Bona-fide boundaries are a result of either spatial discontinuity or qualitative heterogeneity whereas fiat-boundaries are the result of human cognition. Bona-fide boundaries include boundaries of a building, pavements, and roads whereas fiat boundaries include the boundary of a city, district, and block. Here in this research our objective is to create a fiat boundary object at 1:250,000 from bona-fide objects or groups of fiat objects from our source database.

Our approach is based on the notion of *typical* members of a concept for which we want to create the boundary (Tversky 1990). For instance a settlement object will have members such as roads, streets, and trees, but its typical members are the 'buildings'. Similarly for other higher order fiat-objects such as forest, 'trees' are its typical members and for transport networks, roads and railways lines are their typical members. Thus we can hypothesize that these objects in the source database can be used to create the boundary of the fiat object that they are typical members of.

So in order to create a fiat settlement boundary we can select its typical members i.e. building objects for any area from our source database (OS MasterMap Figure 1a). Our Initial research showed that small sparse building groups at the city periphery ('noise') (Figure 1a), tended to make the city boundary overly large. It was necessary to devise a method that identified such noise, in arriving at a boundary definition of city. One approach was to cluster the objects based on proximity that uses a distance threshold 'd_{min}' in order to separate out this periphery noise (Chaudhry and Mackaness 2005). A settlement boundary (city or town) can best be defined by the presence of a large number of buildings in close proximity. But as the example shows (Figure 1b) it can result in some sparsely located buildings falling within the boundary – for example 'soft spikes' that form alongside arterial roads (Figure 1b). On further analysis of the algorithm we observe that the algorithm is forming the boundary based on proximity of buildings whereas it needs to take account of the density of buildings, and to exclude from the region those areas with low densities of buildings or 'noise'.

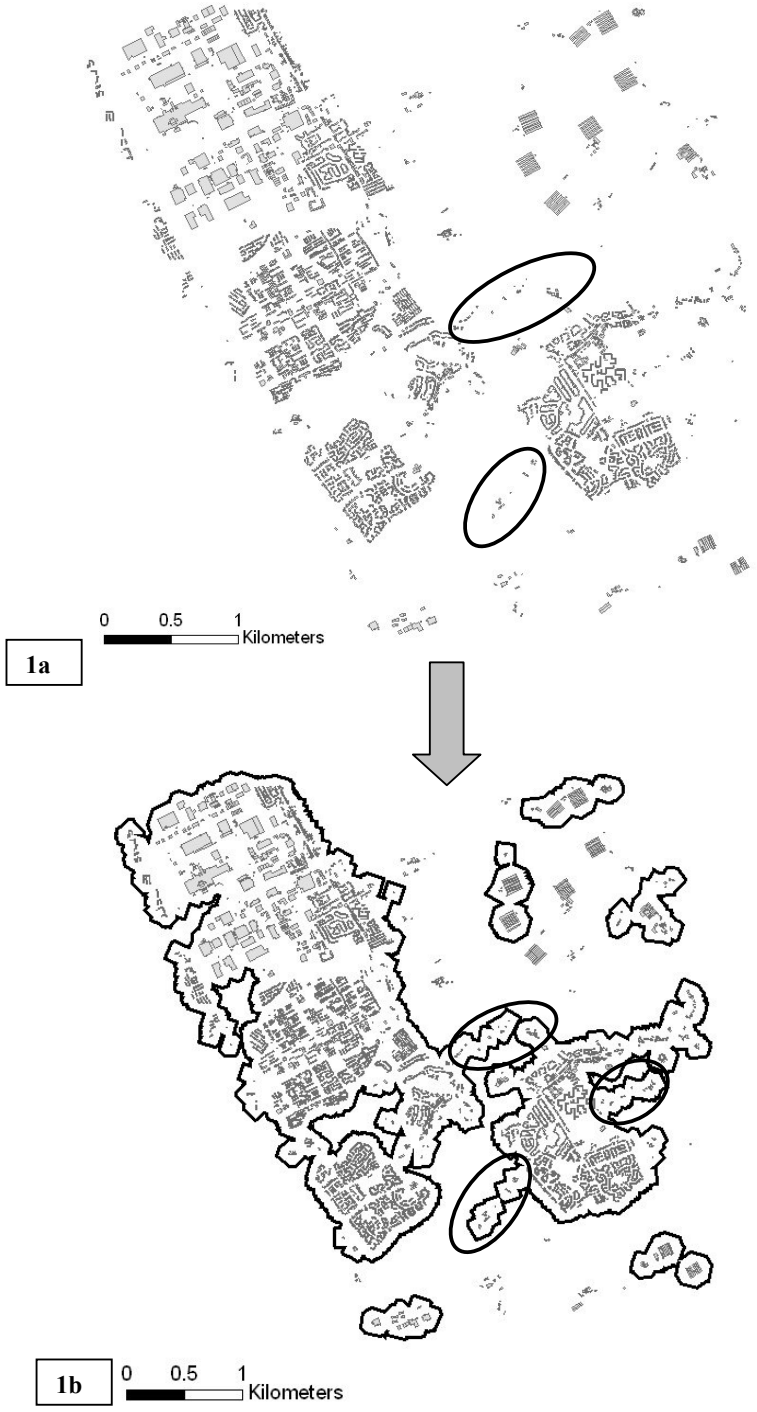


Figure 1: 1a Buildings selected of a random area from OS MasterMap (1:1250/1:10,000). Highlighted area shows noises (low dense buildings). 1b Boundary generated by clustering algorithm(Chaudhry and Mackaness 2005). Some the unwanted soft spikes are ringed.

3.0 Methodology

Here we want to separate objects according to how well they define the higher order object they are part of. One possible solution for identification and separation of these noises highlighted in Figure 1 is by modelling the density. Density is amount per unit size. To model the density we need to partition the space. One way of partitioning the space is to overlay a grid on top of the objects (Wang,

Doihara, and Lu 2002; Doihara, Wang, and Lu 2002) as illustrated in the Figure 2. The principle here is that the amount of area covered by objects in each grid cell is divided by the area of the cell. If it is above a certain threshold the objects are kept otherwise they are not considered. The limitation of this approach is that the amount of excluded objects varies with the orientation and the grid cell size (Figure 3).

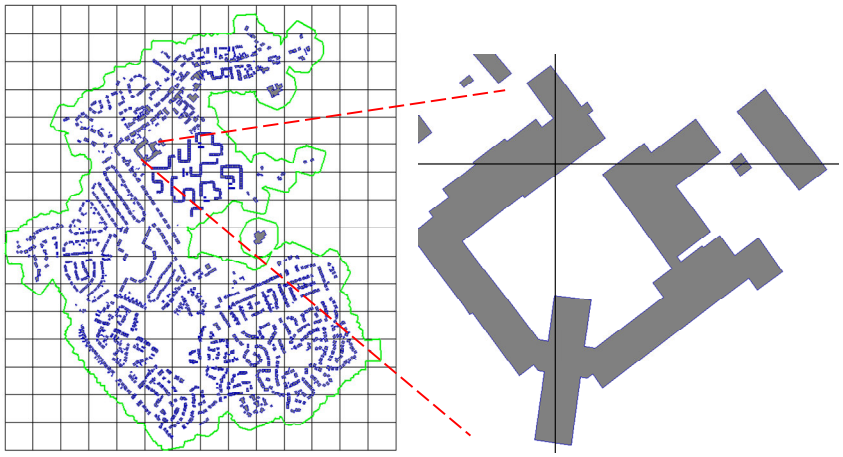


Figure 2: Compartmentalisation of buildings varies with orientation and cell size leading to hard to predict results.

Another way to partition the space is by using the faces created by the road networks of the source database (Chaudhry and Mackaness 2006). A face is a polygon created wherever there is a closed region. Unlike the grids the roads don't run across a building object and their orientation corresponds to that of building objects. The density of objects lying within that face is computed and compared with a threshold as illustrated in Figure 3. The assumption is that there is an even distribution of buildings across the face. In reality, at the city periphery, buildings tend to cluster, with potentially large amounts of open space in the face. It is not possible to separate the buildings from the open space within a face.

Figure 3: Use of road network for identification and removal of noises

3.1 Gravitational Model

The proposed approach is based on the idea of gravitational model. Gravitational pull between two masses is proportional to the size of the mass and the distance between the two masses. If we substitute the building footprint for mass, we can, for any given building, calculate its gravitational 'pull' or attraction. We hypothesise that this attraction should be high for buildings in the

centre of a settlement and will gradually reduce as we approach the edges (since there will be fewer buildings, perhaps smaller, and further apart). We can select a minimum value beneath which buildings are not included in the boundary forming exercise.

Each building is assigned a gravity value – 'g'. The value of g is directly proportional to the area of the object (its area being a surrogate for importance) and the mass is the summation of its neighbours and inversely proportional to the square of the summation of distance (equation1). It is not necessary (nor desirable) to calculate the value of g taking into account all the buildings in the database. From empirical tests it was found that it was only necessary to consider buildings that fell within a 100m radius of the building for which g was being calculated.

$$g_i = \frac{\sqrt{a}_i \sqrt{\sum_{d<=100} a_n}}{d^2} \qquad (1)$$

Where 'g_i' is the gravity value for a given building; 'a_i' is its area; 'a_n' is the area of a proximal buildings; and 'd' limits the zone in which buildings are searched for (100m). Thus the gravitational force 'g_a' will be high for an object whose neighbourhood is made up of dense large objects, and low for objects in more remote parts of the settlement (typically the edge of the settlement). In Figure 2 the objects are grey scaled according to their value of g. An empirically derived threshold value for 'G_{min}' is then used to identify those buildings that do not sufficiently 'belong' to that city.

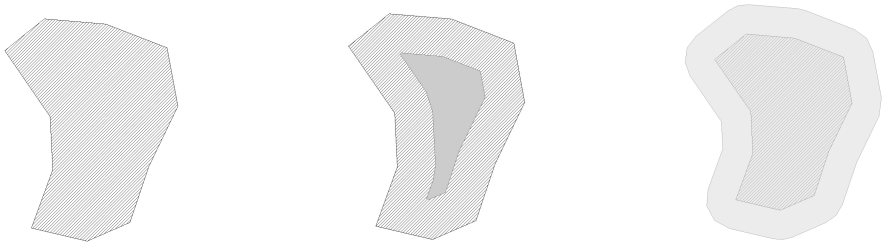
Figure 4: Gravity values for each building (inset shows gradation of values towards margin of city).

3.2 Expansion and Contraction: Boundary Formation

The next step is to generate a boundary. In the first instance objects are buffered according to the following formulae. Objects whose gravity is more then 'G_{min}' are expanded and objects for which the gravity is below 'G_{min}' are contracted. The amount of expansion or contraction is calculated by the following equation.

$$t_i = k\sqrt{g_i} \quad (2) \quad \text{provided } t_i <= k$$

Where 't_i' is the size of expansion or contraction, k is the upper limit of expansion applied to the object with the largest value of 'g'. Here 'k' is a normalisation function. This idea of expansion and contraction is based on the principle of 'richer getting richer' and 'poor getting poorer' (Müller and Wang 1992) as illustrated in Figure 5.

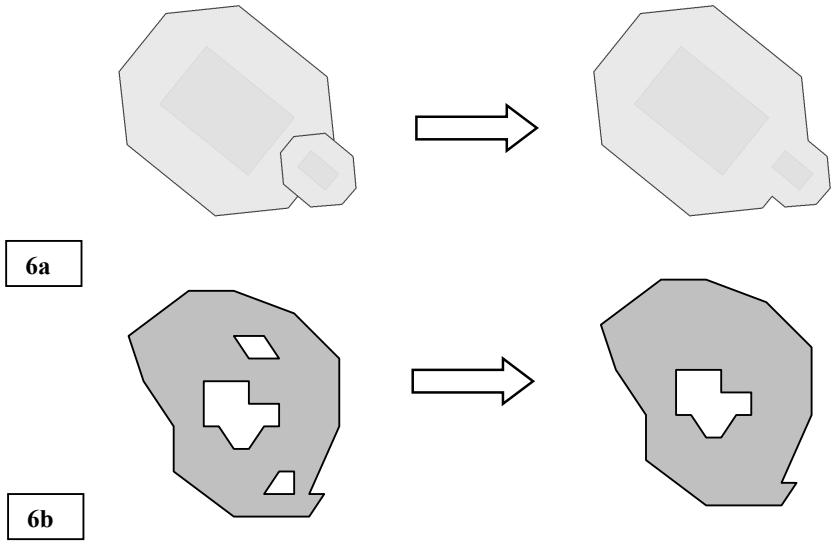


5a **5b** **5c**

Figure 5: 5a Original object. 5b contraction t= -17m 5c expansion t=17m

3.3 Aggregation or Elimination

If boundaries of two or more objects overlap after expansion/contraction their boundaries are aggregated into one boundary polygon (Figure 6a). After aggregation of overlapping boundaries the next step is the selection or elimination of the resultant aggregated object. Here we use the area of the resultant object as a basis for selection. A rule was used derived from an Ordnance Survey map specification: '*For a settlement object the area has to be equal to or greater than 0.01 sq km and for forest it must be equal or greater then 0.25 sq km'(Ordnance 2005)* .Thus small holes or islands (where the area is below the area threshold are absorbed into their containing object (Figure 6b). The idea of elimination of holes and small boundary polygons is based on the principle of generalisation that it should lead to elimination rather then addition of detail.



6a

6b

Figure 6: 6a Overlapping boundaries, result of expansion, are aggregated. 6b The hole/Islands whose area is below area threshold are absorbed into their containing polygon

3.0 Implementation and Results

The methodology presented above is a bottom up approach to creating boundaries. In contrast to a top down approach, where objects from all classes (roads, building, trees, river, and lakes) are considered at the same time, a bottom up approach addresses the generalisation of a particular class at one time. The classification of objects can be based on geometry or on semantics or the combination of the two (Müller and Wang 1992). In this research we have selected objects based on their thematic classification. The overall design of the algorithm is shown in Figure 7.

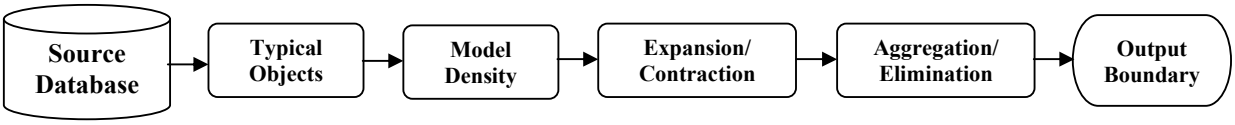


Figure 7: Overall Design:

We applied the above approach on two types of objects (buildings and forest patches) selected from a large scale database (OS MasterMap). The platform selected for the implementation was Java, SQLJ and Oracle 10g. Oracle 10g supports all the geometrical and topological functions defined by OGC as reported in Oosterom et al., (2002).

Taking advantage of the following functionality provided in Oracle (Oracle, 2005):

CREATE INDEX: Creates a spatial index on a column of type SDO_GEOMETRY. The function reduces the execution time.

SDO_GEOM.SDO_DISTANCE: Computes the distance between two geometry objects. The distance between two geometry objects is the distance between the closest pair of points or segments of the two objects.

SDO_GEOM.SDO_BUFFER: Generates a buffer polygon around or inside a geometry object. Used for contraction/expansion.

SDO_AGGR_UNION: Returns a geometry object that is the topological union (OR operation) of the specified geometry objects. Used for aggregation of polygons

SDO_GEOM.SDO_AREA: Returns the area of a two-dimensional polygon.

SDO_UTIL.EXTRACT: Returns the geometry that represents a specified element (and optionally a ring) of the input geometry. Used for removal of holes/islands from a polygon. We also compared the efficiency by implementing the algorithm in JTS (Java Topology Suite) and Jump which are open source platforms. Oracle JTS is much faster since all the spatial objects are created in memory. But this advantage of working in memory reverses when the dataset is large. JTS quickly runs out of memory on normal machines.

Here we present a few results for different input areas selected from OS MasterMap. First the algorithm was applied on building objects (Figure 1a) selected from the source database in order to generate a boundary objects. Figure 8 shows the result obtained from the algorithm. Note that the noises highlighted in Figure 1a which became part of the boundary in our initial clustering algorithm (Figure 1b) are successfully ignored by the algorithm. Figure 9 shows another test data selected and its corresponding settlement boundary objects.

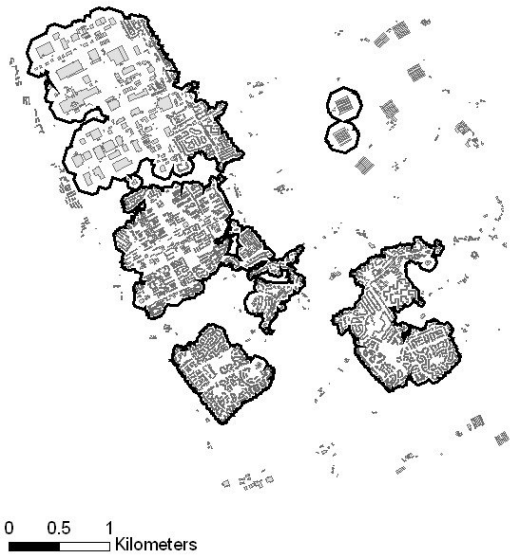


Figure 8: Resultant boundary generated by gravity algorithm and input buildings selected from OS MasterMap. Note the noises highlighted in Figure 1 are no longer part of the resultant boundary. The three settlement boundaries are 'Livingston', 'Mid Calder' and 'East Calder' in Scotland.

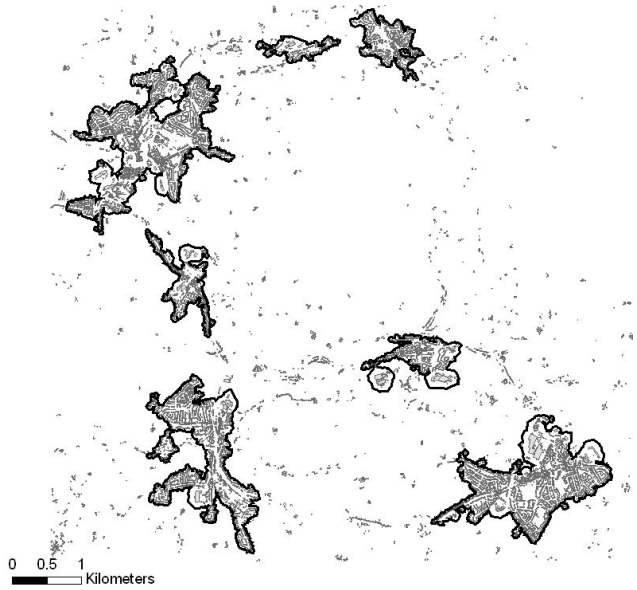


Figure 9: Resultant boundary generated by gravity and input buildings selected from OS MasterMap. The large settlement boundaries are 'New Mills', 'New Town', Furness Vale', ' Chinley', 'Hay Field', 'Chapel –en-le-Frith' and 'Whaley Bridge' in Peak District (England).

As mentioned earlier the above algorithm was also applied on 'Area Patches' (group of trees, lakes). Unlike the building objects where each object represents a particular object these area patches in the source database represents a patch of land which is classified by the surveyor as 'coniferous trees' 'non-coniferous trees' 'scrubs' 'rough grassland' and mixtures of these. For this research we selected area patches which are either coniferous or non-coniferous trees or a mixture of the two (Figure 10). The corresponding forest boundary objects generated by the algorithm are shown in Figure 11.

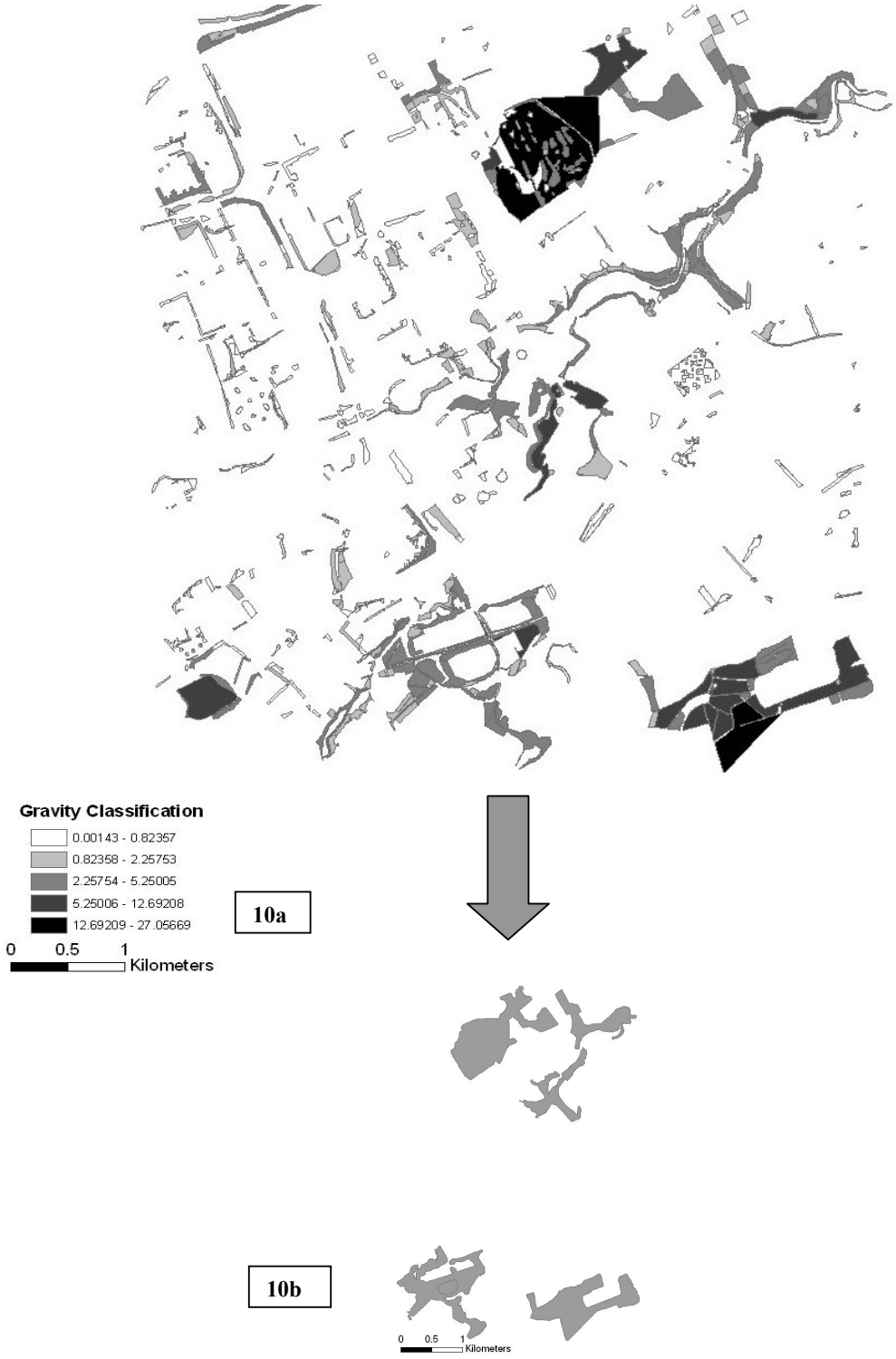


Figure 10: 10a Forest patches selected of random area from OS MasterMap (1:1250/1:10,000). 10b Resultant Boundary generated by gravity algorithm

We also applied the above approach on the dataset (Figure 11a) used by Müller and Wang (1992) for area patch generalisation in order to compare the two approaches. Figure 11b illustrates the gravity classification. Figure 12b and 12 d illustrates resultant boundaries generated by the above algorithm for this dataset using different values of threshold (G_{min}). Figure 11 d is output generated by Muller and Wang (1992). Its important to note unlike

output in Figure 12c that the outputs generated are not cartographic outputs although they can be used as input to cartographic generalisation as explained in subsequent section.

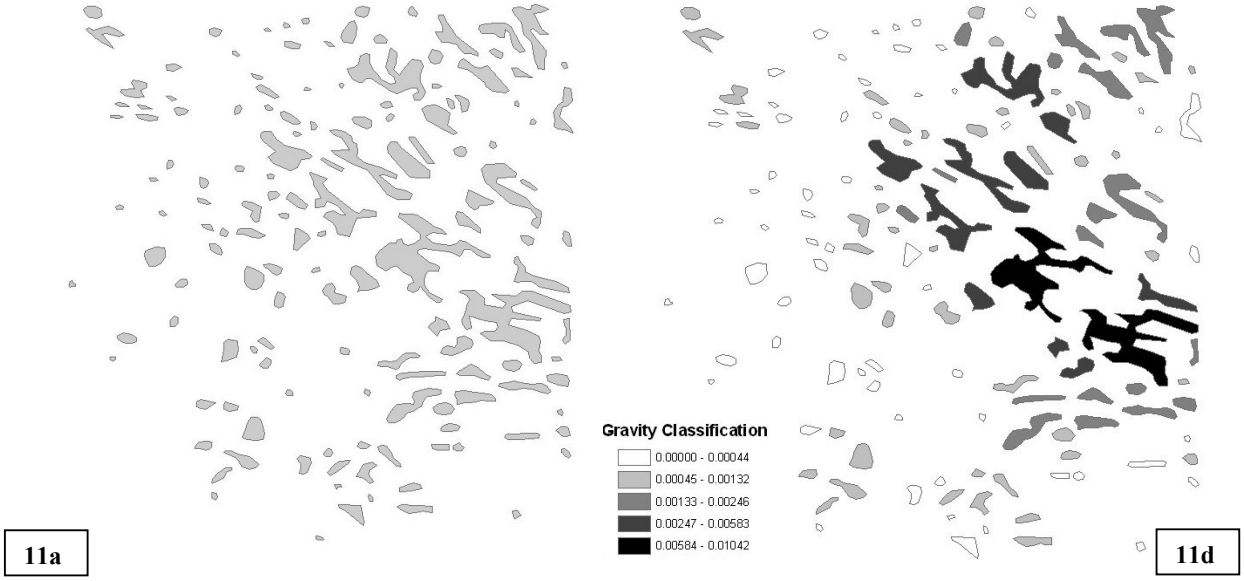


Figure 11: 11a Source data (Müller and Wang 1992). 11b Gravity classification using gravity model

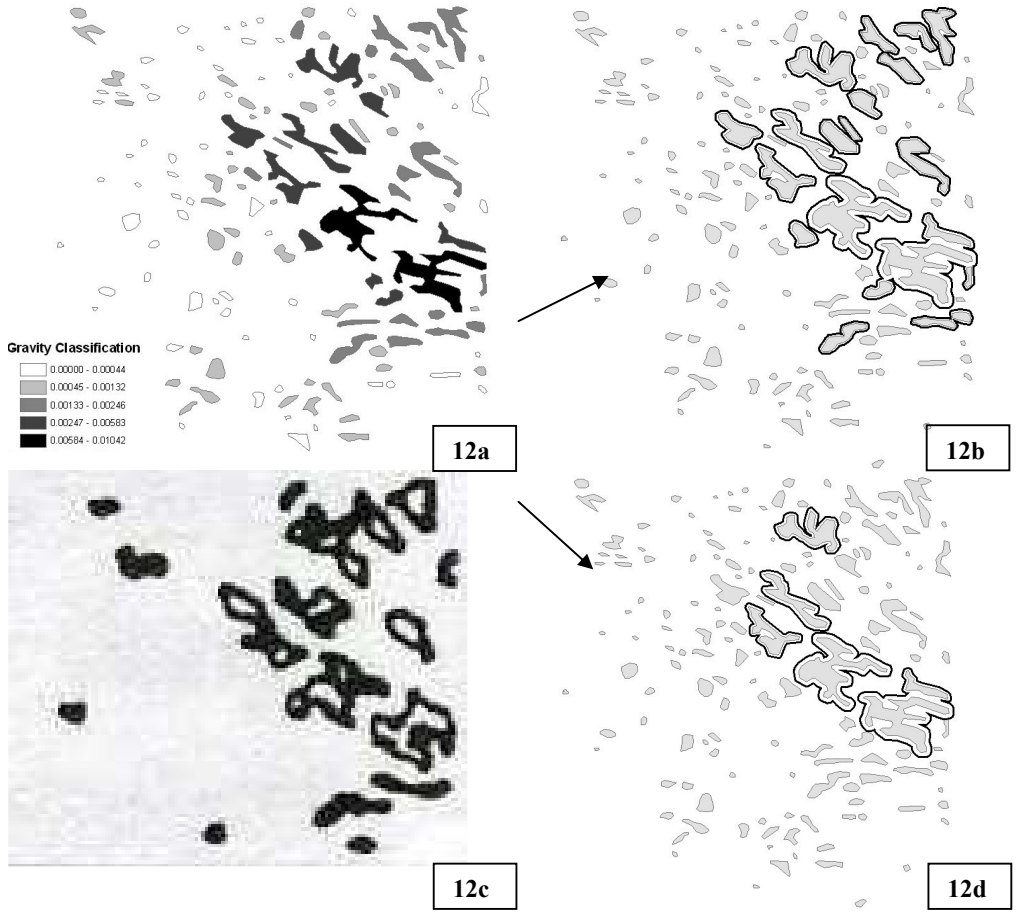


Figure 12: 12a Gravity classification using gravity model presented in this research. 12b Resultant boundary polygons. 12c Output generated by Muller and Wang (1992). 12d Resultant boundary with higher threshold (G_{min}).

5.0 Utility:

The results obtained from the above algorithm can be used as input to cartographic generalisation. So that processes such as displacement, simplification can be applied and a cartographic symbol of the settlement or forest can be generated (say at 1:250,000 scale). A more important utilisation of the above results is there use for extraction of implicit relationships such as the partonomic structures inherent in our source database. This can be achieved by finding the objects in the source database that traverse these boundary objects. The advantage of finding these partonomic relations is firstly that they are useful for object aggregation (Smaalen 2003; Molenaar 1998). Another advantage is that they are useful for performing spatial analysis routines such as finding all objects that are part of a particular city or finding shortest path networks between two cities. According to these partonomic relationships an object's relationship with respect to other objects changes its behaviour and its representational form they are useful for cartographic generalisation. For instance a major road might be modelled in a different way if it is part of a city (servicing the daily commute) as compared to its role in a rural setting – in which the road more serves to connect cities. These different behaviours result in different cartographic visualisations. Since these fiat boundaries are created by vague definitions of the higher order objects (city, mountain, town, forest) there is a degree of fuzziness in the resultant boundary. The boundaries generated by the above algorithm can be used for determining fuzzy membership in terms of 'certain core' and 'certain exterior'(Winter and Thomas 2002). The 'real' boundary will be located somewhere between lower and upper approximation.

6.0 Conclusion

Fiat boundary objects are a result of human conceptualisation of the real world. In this paper we have presented an approach that can be used for the extraction of fiat settlements and fiat forest boundary objects from both bona fide objects and fiat objects. We highlighted the limitations of proximity based clustering algorithm and provided an extension based on gravitational attraction between objects so that low dense 'noises' around the edges can be separated from high dense regions. The approach was applied to objects and area patches. The results obtained compared quite favourably with the manually created urban settlement layer of 1:250,000 map. Further work will look in utilisation of the results for extraction of partonomic relations and for fuzzy membership.

Acknowledgements

We gratefully acknowledge support and funding from the Ordnance Survey and The University of Edinburgh.

References

Casati, R., B. Smith, and A. Varzi. 1998. Ontological Tools for Geographic Representation. In *Formal Ontology in Information Systems*, edited by N. Guarino: IOS Press, Amsterdam.

Chaudhry, O, and W Mackaness. 2005. Visualisation of Settlements Over Large Changes In Scale. Paper read at International Cartographic Association Generalisation Workshop, 7-8 July, at La Coruna.

———. 2006. Density Modelling in Support of Automatic Recognition of Geographical Phenomena in Large Scale Topographic Databases. Paper read at GISRUK, at University of Nottingham.

Doihara, T, P Wang, and W Lu. 2002. An Adaptive Lattice Model and its Application to Map Simplification. Paper read at ISPRS Commission IV, at Ottawa, Canada.

Miller, H.J., and J. Han. 2001. Geographic data mining and knowledge discovery: An overview. In *Geographic Data Mining and Knowledge Discovery.*, edited by H. J. Miller and J. H. Eds: Taylor and Francis London.

Molenaar, M. 1998. *An introduction into the theory of topologic and hierarchical object modeling for geo information systems*: Taylor & Francis.

Müller, J C, and Z Wang. 1992. Area-Patch generalisation: a competitive approach. *The cartographic Journal* 29:137-144.

Ordnance, Survey. 2005. 1:250,000 specifications. *Ordnance Survey internal document*.

Smaalen, J W N van. 2003. Automated Aggregation of Geographic Objects: A New Approach to the Conceptual Generalisation of Geographic Databases. *Doctoral Dissertation, Wageningen University, The Netherlands*.

Smith, B., and D.M Mark. 2003. Do mountains exist? Towards an Ontology of Landforms. *Environment and Planning B:Planning and Design* 30 (3):411-427.

Smith, B., and A. C. Varzi. 1997. The Formal Ontology of Boundaries. *The Electronic Journal of Analytic Philosophy*.

Smith, B., and A. C. Varzi. 2000. Fiat And Bona Fide Boundaries. *Philosophy And Phenomenological Research* 60:401-420.

Tversky, B. 1990. Where partonomies and taxonomies meet In Meanings and Prototypes: Studies on Linguistic Categorization. In *S. Tsohatzidis*: London: Routledge.

Wang, P, T Doihara, and W Lu. 2002. Spatial Generalisation: An Adaptive Lattice Model Based on Spatial Resolution. Paper read at ISPRS Commission IV, at Ottawa, Canada.

Wang W., Yang J., and R. Munoz. 1997. STING: A Statistical Information Grid Approach to Spatial Data Mining. Paper read at 23rd Int. Conf. on Very Large Data Bases, at Athens, Greece.

Winter, S., and B Thomas. 2002. Hierarchical Topological Reasoning with Vague Regions. In *Spatial Data Quality*, edited by S. Winching, P. Fisher and M. F. Godchild: Taylor & Francis.